News Comments: Exploring, Modeling, and Online Prediction

Manos Tsagkias, Wouter Weerkamp, and Maarten de Rijke

ISLA, University of Amsterdam, Science Park 107, 1098 XG Amsterdam, The Netherlands
m.tsagkias@uva.nl, w.weerkamp@uva.nl, derijke@uva.nl

Abstract. Online news agents provide commenting facilities for their readers to express their opinions or sentiments with regards to news stories. The number of user supplied comments on a news article may be indicative of its importance, interestingness, or impact. We explore the news comments space, and compare the log-normal and the negative binomial distributions for modeling comments from various news agents. These estimated models can be used to normalize raw comment counts and enable comparison across different news sites. We also examine the feasibility of online prediction of the number of comments, based on the volume observed shortly after publication. We report on solid performance for predicting news comment volume in the long run, after short observation. This prediction can be useful for identifying news stories with the potential to "take off," and can be used to support front page optimization for news sites.

1 Introduction

As we increasingly live our lives online, huge amounts of content are being generated, and stored in new data types like blogs, discussion forums, mailing lists, commenting facilities, and wikis. In this environment of new data types, online news is an especially interesting type for mining and analysis purposes. Much of what goes on in social media is a response to, or comment on, news events, reflected by the large amount of news-related queries users ask to blog search engines [9]. Tracking news events and their impact as reflected in social media has become an important activity of media analysts [1] and there is a growing body of research on developing algorithms and tools to support this type of analysis (see the related work section below). In this paper, we focus on online news articles plus the comments they generate, and attempt to uncover the factors underlying the commenting behavior on these news articles. We explore the dynamics of user generated comments on news articles, and undertake the challenge to model and predict news article comment volume shortly after publication.

To make things more tangible, consider a striking example of unexpected commenting behavior in response to news stories: March 13, 2009, a busy day for one of the biggest news papers in the Netherlands, *De Telegraaf*. In less than 24 hours, more than 1,500 people commented on *Telegraaf*'s article regarding the latest governmental policy on child benefit abuse. One month later, the Dutch news reported a potential pandemic swine flu, first located in Mexico, but less than five hundred comments were posted to related articles across different news sites, even a week after the first publication. Given that both news events are important to the Dutch society, their numbers of

comments differ greatly. What causes the first story to receive over three times as many comments as the second? What factors contribute to the impact of a news story?

Let us take a step back and ask why we should be interested in commenting behavior and the factors contributing to it in the first place? We briefly mention two types of application for predicting the number of comments shortly after publication. First, in *reputation analysis* one should be able to quickly respond to stories that "take off" and real-time observation and prediction of the impact of news articles is required. Second, the *lay-out decisions* of online news agents often depend on the expected impact of articles, giving more emphasis to articles that are likely to generate more comments, both in their online news papers (e.g., larger headline, picture included) and in their RSS feeds (e.g., placed on top, capitalized).

To come to these applications and answer the questions raised by the example, we need more insight in comments and commenting behavior on online news articles. Our aim is to gain this insight, and use these insights to predict comment volume of news articles shortly after publication. To this end, we seek to answer the following questions:

1. What are the dynamics of user generated comments on news articles? Do they follow a temporal cycle? The answers provide useful features for modeling and predicting news comments.
2. Can we fit a distribution model on the volume of news comments? Modeling the distribution allows for normalizing comment counts across diverse news sources.
3. Does the correlation between number of responses at early time and at later time found in social media such as Digg and Youtube hold for news comments? I.e., are patterns for online responses potentially "universal"? And can we use this to predict the number of comments an article will receive, having seen an initial number?

This paper makes several contributions. First, it explores the dynamics and the temporal cycles of user generated comments in online Dutch media. Second, it provides a model for news comment distribution based on data analysis from eight news sources. And third, it tries to predict comment volume once an initial number of comments is known, using a linear model. In §2 we discuss related work. §3 explores the dataset, and we use insights gained here to try to fit distribution models in §4. Finally, we try to predict comment volume in §5 and conclude in §6.

2 Related Work

Different aspects of the comment space dynamics have been explored in the past. Mishne and Glance [10] looked at weblog comments and revealed their usefulness for improving retrieval and for identifying blog post controversy. Duarte et al. [4] engaged in describing blogosphere access patterns from the blog server point, and identified three groups of blogs using the ratio of posts over comments. Kaltenbrunner et al. [6] measured community response time in terms of comment activity on Slashdot stories, and discovered regular temporal patterns in people's commenting behavior. Lee and Salamatian [7] report that the amount of comments in a discussion thread is inversely proportional to its lifespan after experimenting with clustering threads for two online discussion fora, and for a social networking site. Schuth et al. [12] explore the news comments space of four online Dutch media. They describe the commenters and derive

a method for extracting discussion threads from comments. De Choudhury et al. [3] characterize conversations in online media through their interestingness.

We explore the comment space of online news articles, and model the commenting patterns for multiple news sources. Previous work finds that the distribution of comments over blog posts is governed by Zipf's law [8, 10, 12]. Lee and Salamatian [7] use a Weibull distribution for modeling comments in discussion threads. Kaltenbrunner et al. [5] point to discussions in the literature for selecting the log-normal over the Zipf distribution for modeling; they use four log-normal variants to model response times on Slashdot stories. Ogilvie [11] models the distribution of comment counts in RSS feeds using the negative binomial distribution; a similar approach is taken by Tsagkias et al. [15] to model news comments for prediction prior to publication. Finally, Wu and Huberman [16] find that diggs can be modeled with the log-normal distribution, and Szabó and Huberman [14] model popularity growth of online content using a linear model.

3 Exploring News Comments

In this section we describe our data, comments to online news articles, compare commenting behavior to that in the blogosphere, and discover temporal cycles.

The dataset consists of aggregated content from seven online news agents: *Algemeen Dagblad* (*AD*), *De Pers*, *Financieel Dagblad* (*FD*), *Spits*, *Telegraaf*, *Trouw*, and *WaarMaarRaar* (*WMR*), and one collaborative news platform, *NUjij*. We have chosen to include sources that provide commenting facilities for news stories, but differ in coverage, political views, subject, and type. Six of the selected news agents publish daily newspapers and two, *WMR* and *NUjij*, are present only on the web. *WMR* publishes "oddly-enough" news and *NUjij* is a collaborative news platform, similar to Digg, where people submit links to news stories for others to vote for or initiate discussion. We focus only on the user interaction reflected by user generated comments, but other interaction features may play a role on a user's decision to leave a comment.

For the period November 2008–April 2009 we collected news articles and their comments. Our dataset consists of 290,375 articles, and 1,894,925 comments. The content is mainly written in Dutch. However, since our approach is language independent and we believe that the observed patterns and lessons learned apply to news comments in other countries, we could apply our approach to other languages as well.

3.1 News comments vs. blog post comments

The commenting feature in online news is inspired by the possibility for blog readers to leave behind their comments. Here, we look at general statistics of our news sources and comments, and compare these to commenting statistics in blogs as reported in [10]; the numerical summary can be found in Table 1. News comments are found to follow trends similar to blog post comments. The total number of comments is an order of magnitude larger than the total number of articles, which is positively correlated with the case of influential blogs. In general, about 15% of the blog posts in the dataset in [10] receives comments, a number that increases for the news domain: the average percentage of commented articles across all sources in our dataset is 23%. *Spits* and *WMR* display the interesting characteristic of receiving comments on almost every article they publish. This can be explained by the two sites having very simple commenting facilities. In contrast, *Trouw* has the lowest ratio of commented articles: commenting is enabled

News agent	Total articles (commented)	Total comments	Comments per article w/ comments mean	median	st.dev	Time (hrs) 0–1 com.	1–last com.
AD	41 740 (40%)	90 084	5.5	3	5.0	9.4	4.6
De Pers	61 079 (27%)	80 72	5.0	2	7.5	5.9	8.4
FD	9 911 (15%)	4 413	3.0	2	3.8	10.	9.3
NUjij	94 983 (43%)	602 144	14.7	5	32.3	3.1	6.3
Spits	9 281 (96%)	427 268	47.7	38	44.7	1.1	13.7
Telegraaaf	40 287 (21%)	584 191	69.9	37	101.6	2.5	30.2
Trouw	30 652 (8%)	19 339	7.9	4	10.3	11.7	8.1
WMR	2 442 (100%)	86 762	35.6	34	13.08	1.1	54.2

Table 1. Dataset statistics of seven online news agents, and one collaborative news platform (*NUjij*) for the period November 2008–April 2009.

only for some of the articles, partially explaining the low ratio of commented articles. Another reason can be the content's nature: WMR's oddly-enough news items are more accessible and require less understanding increasing the chance to be commented.

Half of the news sources receive the same number of comments as blogs (mean 6.3), whereas the other half enjoys an order of magnitude more comments than blogs. Looking at reaction time, the time required for readers to leave a comment, it is on average slower for news ($\sim$ 6 hrs) than for blogs ($\sim$ 2 hrs), although this differs significantly per news source. A speculation on the reason underlying these differences can be the news source's readers demographics, e.g., tech savvies or youngsters are rather quick to react, whilst older people, less acquainted with the internet, access the online version of the news papers less frequently.

3.2 Temporal cycles of news comments

We perform an exploration of temporal cycles governing the news comment space. We look at three levels of temporal granularity: monthly, daily, and hourly. In our dataset, the volume of comments ranges two orders of magnitude making the comparison of raw comment counts difficult. We therefore report comments in z-scores: z-scores represent how many σ's (standard deviations) the score differs from the mean, and allows for comparison across sources.

Looking at comment volume per month in Fig. 1, we observe months with high and low comment volume, either reflecting the importance of published news, or the seasonal user behavior. For example, March shows the highest comment volume across the board, and November shows the least for most sources.

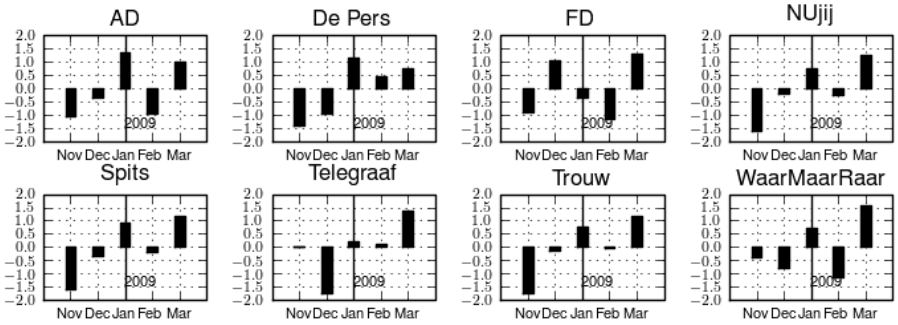

Fig. 1. Comments per month and per source. Vertical line is a year separator.

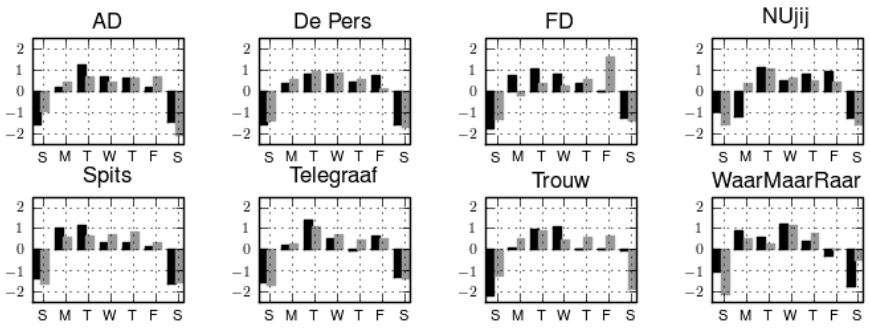

Fig. 2. Comments (black) and articles (grey) per day of the week and per source.

We explore the comment volume per day of the week in Fig. 2: weekdays receive more comments compared to weekends, with Wednesday being, on average, the most active day and Sunday the least active day across the board. These results are in agreement with the activity observed in social networks such as Delicious, Digg, and Reddit.[1] Comparing the number of comments to the number of articles published per day, most sources show an insignificant, negative correlation ($p \gg 0.05$). Three sources, however, have articles and comments highly correlated, but differ in polarity: *FD* and *Trouw* show a negative correlation and *NUjij* shows a positive correlation. The variety found in correlation polarity likely indicate the commenting behavior of a source's audience.

Finally, we look at the distribution of comments throughout the day. Fig. 3 reveals a correlation between posted comments, sleep and awake time, as well as working, lunch and dinner time. The comment volume peaks around noon, starts decreasing in the afternoon, and becomes minimal late at night. Interesting exceptions are *NUjij*, the collaborative news platform, and *FD*, a financial newspaper: comment volume in *NUjij* matches with blog post publishing [8], which has a slow start and gradually peaks late in the evening. *FD* on the other hand receives most of its comments early in the morning, and then drops quickly. This is in line with the business oriented audience of *FD*.

Overall, the commenting statistics in online news sources show similarities to those in the blogosphere, but are nevertheless inherent characteristics of each news source. The same goes for the temporal cycles, where we see similar patterns for most sources,

[1] http://3.rdrail.net/blog/thurday-at-noon-is-the-best-time-post-and-be-noticed-pst/

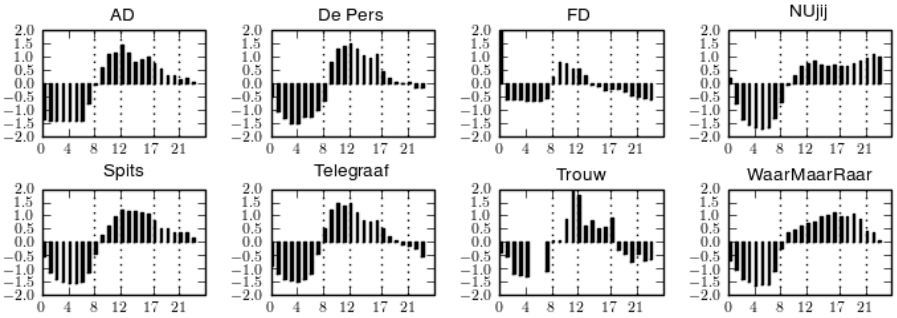

Fig. 3. Comments per hour and per source.

but also striking differences. These differences in general and temporal characteristics possibly reflect the credibility of the news organisation, the interactive features they provide on their web sites, and their readers' demographics [2].

4 Modeling News Comments

In this section we seek to identify models (i.e., distributions) that underly the volume of comments per news source. We do so (1) to understand our data, and (2) to define "volume" across sources. If two articles from two sources receive the same number of comments, do they expose the same volume? Ten comments may signify a high volume for an article in one source, but a low volume in another. Expressing comment volume as a normalized score offers a common ground for comparing and analyzing articles between sources. Our approach is to express a news article's comment volume as the probability for an article from a news source to receive x many comments. We consider two types of distribution to model comment volume: log-normal and negative binomial.

Recall that the log-normal distribution is a continuous distribution, with probability density function defined for $x > 0$, cf. (1), and the two parameters μ (the mean) and σ (the standard deviation of the variable's natural logarithm) affect the distribution's shape. For a given source we estimate the parameters using maximum likelihood estimation.

$$LN_{pdf}(x; \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(lnx-\mu)^2}{2\sigma^2}} \tag{1}$$

The negative binomial distribution is a discrete distribution with probability mass function defined for $x \geq 0$, with two parameters r ($r-1$ is the number of times an outcome occurs) and p (the probability of observing the desired outcome), cf. (2). There is no analytical solution for estimating p and r, but they can be estimated numerically.

$$BN_{pmf}(k; r, p) = \binom{k+r-1}{r-1} p^r (1-p)^k \tag{2}$$

For evaluating the models' goodness of fit we choose the χ^2 test. χ^2 is a good alternative to the widely used Kolmogorov-Smirnov goodness of fit test due to its applicability to both continuous and discrete distributions [13]. The metric tests whether a sample of observed data belongs to a population with a specific distribution. Note that, the test requires binned data, and as such is sensitive to the number of chosen bins.

For each news source we estimate the parameters for the log-normal and the negative binomial distributions over the entire period of our dataset (see Fig. 4), and report χ^2 goodness of fit results in Table 2. Both distributions fit our dataset well, with low χ^2 scores denoting strong belief that the underlying distribution of the data matches that of log-normal and negative binomial. Log-normal is rejected for *WaarMaarRaar* possibly because it failed to reach close enough the peak observed at 25 comments. We stress that the results should be taken as indicative, mainly due to the sensitivity of χ^2 to the number of bins (here 30). We experimented with different bin sizes, and observed that for different number of bins either the log-normal, or the negative-binomial failed to describe all sources. Although searching for the optimal number of bins for both distributions to fit all sources could be interesting, we did not exhaust the entire potential. An example of the test's sensitivity is shown in Table 3 where log-normal displays very similar results to negative-binomial even for the source that failed the χ^2 test.

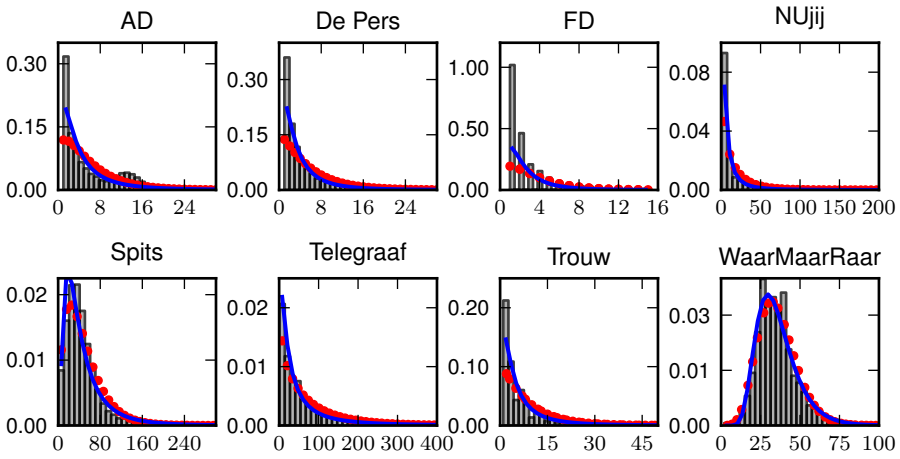

Fig. 4. Modeling comment volume distribution per source using the continuous log-normal (blue line), and the discrete negative binomial distribution (red dots). Grey bars represent observed data. Probability density is on y-axis, and number of comments (binned) is on x-axis.

The final decision on which distribution to favor, depends on the data to be modeled and task at hand. From a theoretical point of view, negative binomial seem better suited to the task of modeling comments: comments are not a continuous but discrete variable. From a practical point of view, for the same task, log-normal parameters are less expensive to estimate and the results match closely those of negative binomial.

The results of our data exploration and modeling efforts are put to the test in the next section, in which we explore the correlation between comment volume shortly and longer after publication.

5 Predicting Comment Volume After Publication

Predicting the number of news comments *prior* to publication in the long term has proved to be very challenging [15]. Szabó and Huberman [14] published promising

News site	Log-normal		Negative binomial	
	χ^2 score	p-value	χ^2 score	p-value
AD	0.08	1.00	0.08	1.00
De Pers	0.59	1.00	0.64	1.00
FD	0.18	1.00	0.26	1.00
NUjij	0.06	1.00	0.06	1.00
Spits	0.67	1.00	1.42	1.00
Telegraaf	0.04	1.00	0.04	1.00
Trouw	0.56	1.00	0.98	1.00
WaarMaarRaar	**236.89**	0.00	0.15	1.00

Table 2. χ^2 goodness of fit for log-normal and negative binomial distributions at 0.10 significance level. Boldface indicates rejection of the null hypothesis: observed and expected data belong to the same distribution.

Distribution	Comments for ICDF @ 0.5							
	AD	*De Pers*	*FD*	*NUjij*	*Spits*	*Telegraaf*	*Trouw*	*WMR*
Log-normal (LN)	3	3	2	6	36	32	4	34
Negative binomial (NB)	3	3	1	8	39	43	5	33

Table 3. Number of comments, per source corresponding at 0.5 of the inverse cumulative distribution function (ICDF).

work on predicting the long term popularity of Digg stories (measured in diggs), and Youtube videos (measured in views) *after* observing how their popularity evolves in the first hours of publication. First, we are interested in finding out whether the correlation between early and late popularity found by Szabó and Huberman also holds for the news comments space. Then, assuming such a relation has been confirmed, it can be employed for predicting the comment volume of a news story.

We begin to explore the relation between early and late comment volume by looking at the similarities of news comments and other online user generated content. In Section 3.2 we reported on the circadian pattern underlying news comment generation, which is found to be similar to blog posts [10], Diggs and Youtube video views [14]. The existence of a circadian pattern implies that a story's comment volume depends on the publication time, and therefore not all stories share the same prospect of being commented; stories published during daytime—when people comment the most—have a higher prior probability of receiving a comment.

Taking into account the above, publication time adds another dimension of complexity in finding temporal correlations. To simplify our task, we introduce a temporal transformation from real-time to *source-time*, following [14], a function of the comment volume entering a news site within a certain time unit. I.e., *source-time* is defined as the time required for $\overline{x}_i$ comments to enter a news agent system i, where $\overline{x}_i$ stands for the average comments per hour cast to a particular source, and is the division of a source's total comments by the total number of hours that we have data for. Consequently, *source-time* has the property of expanding or condensing the real-time scale in order to keep the ratio of incoming comments per hour fixed. Once the number of comments per time unit has been fixed, all stories share the same probability to be commented independently of their publication time. In the rest of this section, story comments are measured in their news agent specific *source-time*, e.g., for *Trouw* we measure in *trouw-time*, for *WMR* in *wmr-time*, etc. Once the temporal transformation is in place, we need a definition for *early* and *late* time, between which we are interested in discovering a correlation. We introduce *reference time* t_r as "late" time, and we set it at 30 source-days after the story has been published. For "early" time, we define *indicator time* t_r to range from 0 to t_r in hourly intervals [14]. Some news agents disable comments after a certain period. As a result, there are articles that constantly reach their maximum comments before t_r, however we have not marked them separately.

We choose Pearson's correlation coefficient ρ to measure the correlation strength between reference and indicator times. Using articles with more than one comment, we compute ρ in hourly intervals from publication time to reference time for all sources over the entire period of the dataset. Fig. 5 shows that the number of comments per source increases exponentially, yet with different rates, reflecting the commenting rules of each site: the time a story remains visible on the front page, for how long comments

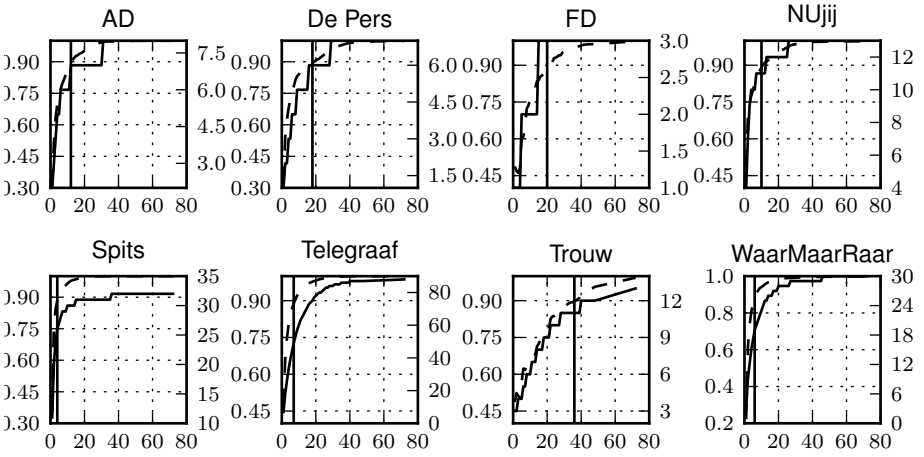

Fig. 5. Comment counts averaged over all stories (right y-axis, solid line), and ρ between indicator, and reference time (left y-axis, dashed line). Indicator time shown at x-axis. Vertical line shows the indicator time with $\rho \geq 0.9$.

are enabled, etc. In the same figure we show a positive correlation that grows stronger as t_i approaches t_r due to stories that saturate to their maximum number of comments. The curve slope indicates how fast stories reach their maximum number of comments, e.g., *Spits* displays a very steep comment volume curve meaning that most stories stop receiving comments short after publication. Looking at when sources reach strong correlation ($\rho > 0.9$) we find that the corresponding indicator times reflect the average commenting lifespan of each source (see Table 1). In contrast to our expectations that *NUjij*, the collaborative news platform, follows a fast correlation pattern similar to Digg (0.98 after the 5th digg-hour), our findings suggest that a strong correlation is achieved much later (ρ at 0.90 after 11 source-hours). Although, nujij-time and digg-time are not directly comparable, we can compare the average user submissions entering each system per hour: 5.478 diggs vs. 140 comments. The difference in order of magnitude can be explained by the different popularity levels enjoyed by the two sites. One could argue that digg-ing and commenting are different tasks: on the one hand, commenting, similarly to writing, asks for some reflection on how to verbalize one's thoughts regardless of the size or the quality of the comment. On the other hand, digg-ing requires the click of a button, rendering the task easier, and hence more attractive to participate.

Given the exponential accumulation of comments over time, a logarithmic scale is appropriate for plotting. In contrast to Diggs or YouTube views, comments do not scale more than two orders of magnitude (compare $10^0 - 10^2$ for comments to $10^1 - 10^4$ for Diggs and Youtube views). Despite the difference in scale, our data shows an emerging pattern similar to Youtube, where a bump is observed in the middle range of early comments. From Fig. 6 two groups of stories emerge, both resulting in many comments: one with stories that begin with too few comments in early indicator times, and one with stories that begin with many comments. This pattern is different from Digg or Youtube where a linear correlation is evident in similar graphs [14].

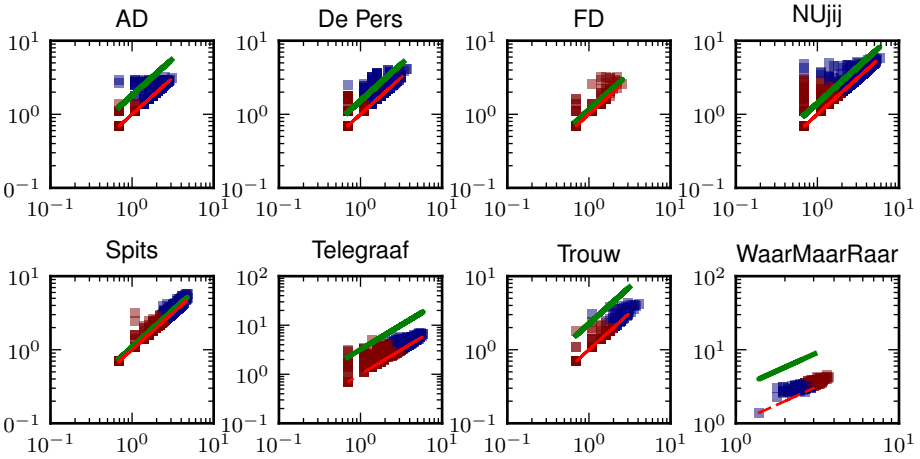

Fig. 6. Correlation of news stories comment volume per source between 2 hours, and 30 days after publication. Number of comments at $t_i(2)$ is x-axis, and comments at t_r is y-axis. K-means separates stories in two clusters depending on their initial comments. Green line shows a fitted model using only the upper stories, with slope fixed at 1. Red dashed line marks the boundary where no stories can fall below.

For our prediction experiments, we are interested in minimizing noise to improve performance, and hence could exploit the emerging clusters by eliminating stories with too few comments at early indicator times. Since these stories exhibit a rather random pattern with regards to their final number of comments, we employ k-means clustering in an attempt to separate them from stories that show a more consistent pattern.

We follow [14] and estimate a linear model on a logarithmic scale for each source in our dataset. The linear scale estimate $\hat{N}_s$ for a story s at indicator time t_i given t_r is defined as $\hat{N}_s(t_i, t_r) = exp[ln(\alpha_0 N_s(t_i)) + \beta_0(t_i) + \sigma^2/2]$, where $N_s(t_i)$ is the observed comment counts, α_0 is the slope, β_0 is the intercept, and σ^2 is the variance of the residuals from the parameter estimation.

For evaluating our model we choose the relative squared error metric averaged over all stories from a certain source at t_i given t_r.

$$QRE(s, t_i, t_r) = \sum_c \left[\frac{\hat{N}_s(t_i, t_r) - N_s(t_r)}{N_s(t_r)} \right]^2 \tag{3}$$

For our experiments, we split our dataset in training and testing for each source. The training sets span from November 2008—January 2009, and the test sets cover February 2009. Model parameters are estimated on the training set, and QREs are calculated on the test set using the fitted models.

We define three experimental conditions based on which we estimate model parameters using our training set: (M1) using in the upper end stories as clustered by k-means, and fixing the slope at 1, (M2) using all stories, and fixing the slope at 1, and (M3) using all stories. Fig. 7 illustrates QREs for the three experimental conditions up to 25 hours after observation; we choose not to include all indicator times up to reference

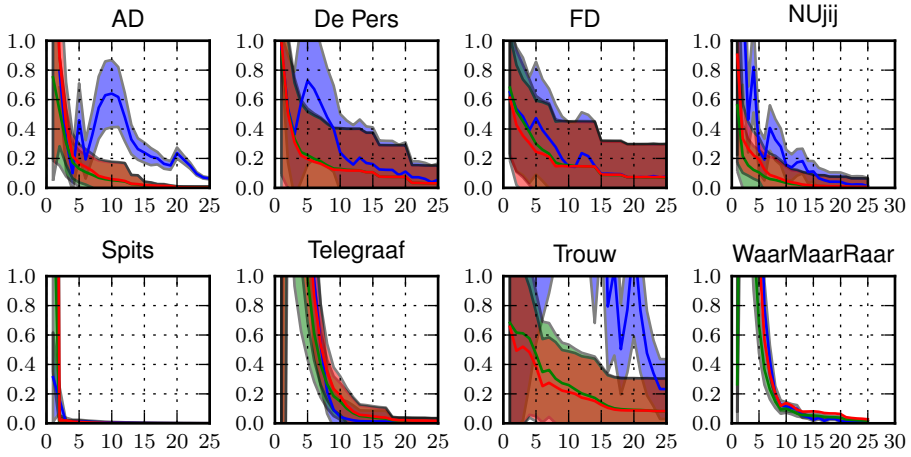

Fig. 7. Relative square error using Model 1 (blue line), Model 2 (green line), and Model 3 (red line). Standard deviation is shown in the shaded areas around the lines. QRE on y-axis, indicator time on x-axis

time to increase readability of the details at early times. From the three experimental conditions, M1 proved to underperform in most cases. M2 and M3 demonstrate similar performance across the board with one slightly outperforming the other depending on the source. QREs decrease to 0 as we move to reference time, followed by a similar decrease in standard error. M2 demonstrates strong predictive performance indicated by low QRE < 0.2 for all sources, in less than 10 hours of observation. The QREs converge to 0 faster for some sources and slower for others, exposing the underlying commenting dynamics of each source as discussed earlier.

In this section we looked at natural patterns emerging from news comments, such as the possible correlation of comment counts on news stories between early and later publication time. A relation similar to the one observed for Digg and Youtube has been confirmed, allowing us to predict long term comment volume with very small error. We observed that different news sources ask for different observation times before a robust prediction can be made. Using QRE curves one can find the optimum observation time per source, that balances between short observation period and low error.

6 Conclusion and Outlook

We studied the news comments space from seven online news agents, and one collaborative news platform. Commenting behavior in the news comments space follows similar trends as the behavior in the blogosphere. Our news sources show quite similar temporal cycles and commenting behavior, but that mainly the differences herein reflect differences in readers' demographics and could prove useful in future research.

As to modeling comments from various news agents, we compared the log-normal and negative binomial distributions. These estimated models can be used to normalize raw comment counts and enable comparison, and processing of articles from different news sites. According to χ^2 goodness of fit test, the underlying distribution of news comments matches with either log-normal or negative binomial. The latter is a discrete

distribution and suits the task better, yet in our our setup log-normal showed similar results and parameter estimation for log-normal is computationally less expensive.

Finally, we looked at the feasibility of predicting the number of comments at a late time, based on the number of comments shortly after publication. Our goal was to find patterns similar to other online content such as Digg, and Youtube. We confirmed this relation, and exploited its potential using linear models. Our results showed that prediction of the long term comment volume is possible with small error after 10 source-hours observation. This prediction can be useful for identifying news stories with the potential to "take off," and can for example be used to support front page optimization for news sites.

Acknowledgments. This research was supported by the DAESO and DuOMAn projects carried out within the STEVIN programme which is funded by the Dutch and Flemish Governments under project numbers STE-05-24 and STE-09-12, and by the Netherlands Organisation for Scientific Research (NWO) under project numbers 640.001.501, 640.002.501, 612.066.512, 612.061.814, 612.061.815, 640.004.802.

References

[1] D. L. Altheide. *Qualitative Media Analysis (Qualitative Research Methods)*. Sage Pubn Inc, 1996.

[2] D. S. Chung. Interactive features of online newspapers: Identifying patterns and predicting use of engaged readers. *J. Computer-Mediated Communication*, 13(3):658–679, 2008.

[3] M. De Choudhury, H. Sundaram, A. John, and D. D. Seligmann. What makes conversations interesting? In *WWW'09*, pages 331–331, April 2009.

[4] F. Duarte, B. Mattos, B. A., A. V., and A. J. Traffic characteristics and communication patterns in blogosphere. In *ICWSM'06*, March 2007.

[5] A. Kaltenbrunner, V. Gomez, and V. Lopez. Description and prediction of slashdot activity. In *LA-WEB'07*, pages 57–66, 2007.

[6] A. Kaltenbrunner, V. Gómez, A. Moghnieh, R. Meza, J. Blat, and V. López. Homogeneous temporal activity patterns in a large online communication space. *CoRR*, 2007.

[7] J. G. Lee and K. Salamatian. Understanding the characteristics of online commenting. In *CONEXT'08*, pages 1–2, 2008.

[8] G. Mishne and M. de Rijke. Capturing global mood levels using blog posts. In *AAAI-CAAW*, pages 145–152, 2006.

[9] G. Mishne and M. de Rijke. A study of blog search. In *ECIR'06*. Springer, April 2006.

[10] G. Mishne and N. Glance. Leave a reply: An analysis of weblog comments. In *Third annual workshop on the Weblogging ecosystem*, 2006.

[11] P. Ogilvie. Modeling blog post comment counts, July 2008. URL http://livewebir.com/blog/2008/07/modeling-blog-post-comment-counts/.

[12] A. Schuth, M. Marx, and M. de Rijke. Extracting the discussion structure in comments on news-articles. In *WIDM'07*, pages 97–104, 2007.

[13] D. J. Sheskin. *Handbook of Parametric and Nonparametric Statistical Procedures*. Chapman & Hall/CRC, USA, 2000.

[14] G. Szabó and B. A. Huberman. Predicting the popularity of online content. *CoRR*, abs/0811.0405, 2008.

[15] M. Tsagkias, W. Weerkamp, and M. de Rijke. Predicting the Volume of Comments on Online News Stories. In *CIKM'09*, pages 1765–1768, 2009.

[16] F. Wu and B. A. Huberman. Novelty and collective attention pnas. URL http://www.pnas.org/content/104/45/17599.abstract.